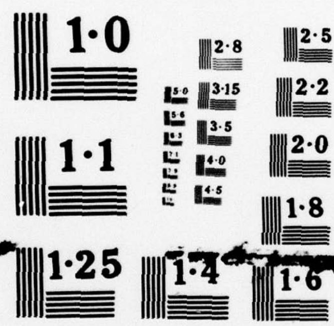
AD-A068 205 TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS F/G 12/1
NONPARAMETRIC TESTS OF INDEPENDENCE.(U)
FEB 79 J CARMICHAEL DAAG29-78-G-0180
UNCLASSIFIED TR-A-1 ARO-16228.2-M NL

1 OF 1
ADA
068205

END
DATE
FILMED
6-79
DDC



NATIONAL BUREAU OF STANDARDS
MICROCOPY RESOLUTION TEST CHART

[redacted]

12 18 ARO 19 16228.2-M

TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843

INSTITUTE OF STATISTICS
Phone 713 - 845-3141

NONPARAMETRIC TESTS OF INDEPENDENCE

Jean-Pierre Carmichael
Université Laval, Québec, Canada

Technical Report No. A-1

February 1979

Texas A & M Research Foundation
Project No. 3861

"Maximum Robust Likelihood Estimation and Non-parametric Statistical Data Modeling"

Sponsored by the U.S. Army Research Office

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: Technical Report No. A-1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): 6 Nonparametric Tests of Independence.		5. TYPE OF REPORT & PERIOD COVERED: 9 Technical rept.
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): 10 Jean-Pierre Carmichael		8. CONTRACT OR GRANT NUMBER(s): 15 DAAC29-78-G-0180
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Texas A&M University, Institute of Statistics, College Station, TX 77843		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Army Research Office, Research Triangle Park, NC 27709		12. REPORT DATE: 11 February 1979
		13. NUMBER OF PAGES: 12
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): 12 7p.		15. SECURITY CLASS. (of this report): Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA 14 TR-A-1

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Nonparametric tests, nonparametric regression, bivariate dependence, statistical data modelling

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We present in this report several tests of independence that have their roots in the theoretical framework developed by Parzen (1977) for nonparametric regression.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

347 380

Gu

NONPARAMETRIC TESTS OF INDEPENDENCE

by

Jean-Pierre Carmichael

Introduction

We present in this report several tests of independence that have their roots in the theoretical framework developed by Parzen (1977) for nonparametric regression. For the sake of completeness, we repeat the argument.

Let (X, Y) be random variables with joint distribution function $F_{X,Y}(\cdot, \cdot)$ and joint density $f_{X,Y}(\cdot, \cdot)$. Let $U_1 = F_X(X)$ and $U_2 = F_Y(Y)$, then the joint distribution of U_1 and U_2 is

$$D(u_1, u_2) = F_{X,Y}\left(Q_X(u_1), Q_Y(u_2)\right)$$

and the joint density is

$$d(u_1, u_2) = \frac{f_{X,Y}\left(Q_X(u_1), Q_Y(u_2)\right)}{f_X\left(Q_X(u_1)\right) \cdot f_Y\left(Q_Y(u_2)\right)} .$$

$Q_X(\cdot)$ is the quantile function of X, $f_X(\cdot)$ is its marginal density and $F_X(\cdot)$ is its marginal distribution.

Parzen called $d(\cdot, \cdot)$ the regression-density of (X, Y) because

$$E[Y | X = Q_X(u_1)] = \int_0^1 Q_Y(u_2)\, d(u_1, u_2)\, du_2 .$$

The hypothesis of independence of X and Y can be expressed in terms of $D(\cdot)$ and $d(\cdot)$:

X and Y are independent if and only if

H_1 $\quad d(u_1, u_2) \equiv 1$

or

H_2 $\quad D(u_1, u_2) = u_1 \cdot u_2$

It is customary to have in mind some alternative when proposing a test of hypothesis. In the case of tests of independence, the alternative is rarely dependence because this concept is too broad (except in the bivariate normal case).

1. Rank Plots

Given observations $\{(Xi, Yi)\}_{i=1}^n$ from a population with distribution function $F_{X,Y}(\cdot, \cdot)$, we consider the following transformation

$$(Xi,\ Yi) \rightarrow \left(\tilde{F}_X(Xi),\ \tilde{F}_Y(Yi)\right)$$

where $\tilde{F}_X(\cdot)$ is the empirical distribution function of the X-component. We order the resulting pairs on the first component to obtain points of the form $(i/n,\ Ri/n)$, where Ri is the rank of the concomitant of the i^{th} ordered X, e.g., if $X_{(1)} \le \ldots \le X_{(\ )}$, Ri is the rank of the Y-variate associated with $X_{(i)}$.

The points $\{(\frac{i-1}{n},\ \frac{Ri-1}{n})\}_{i=1}^{n}$ are the points of mass of the empirical bivariate distribution function $\tilde{D}(\cdot,\cdot)$. The rank plot is simply the scattergram of these points. By counting how many points of the scattergram are included in the rectangle $[0,u_1]\times[0,u_2]$ and dividing by n, one obtains the estimate $\tilde{D}(u_1,u_2)$ that could be compared to the null value of $u_1 \cdot u_2$. The problem is that the hypothesis of independence says that $D(u_1,u_2) = u_1 u_2$ for all (u_1, u_2).

If we look at the scattergram itself, the hypothesis of independence says that the unit square should be filled uniformly. Visual inspection can detect patterns and clusters and should be performed as a first step, even though no level of significance can be attached directly to that operation.

The rank plot also gives indication on the behavior of the regression of Y on X, (e.g., monotonicity, cycles).

2. Concomitant Plots

A second possible transformation is

$$(Xi,\ Yi) \rightarrow (i/n,\ Y_{Ri}) .$$

Then, $\{Y_{Ri}\}_{i=1}^{n}$ is a sample from a time series, with observations taken at equidistant points of the form $\{i/n\}_{i=1}^{n}$.

Under the null hypothesis of independence, this sample would come from a "white noise" time series. This can be tested using, among others, Parzen's CAT criterion or Akaike's criterion, etc.

Another possibility is to use $\Phi^{-1}\left(\frac{Ri + 1/2}{n}\right)$ instead of Y_{Ri}, where $\Phi^{-1}(\cdot)$ is the normal quantile function, as we did to produce Table 1.

Table 1

% Correct Decisions based on 100 Replications using CAT criterion.

ρ	N = 20	40	100
0.0	77	79	73
0.1	16	20	24*
0.2	17	21	36*
0.3	19	22	36*
0.4	20	37	74*
0.5	25	55	96*
0.6	36	73	98*
0.7	61	96	100*
0.8	80	100	100*
0.9	95	100	100*

* based on 50 replications only.

The scattergram of these points is what we call the concomitant plot. Visual inspection can help us *form* an opinion about the data. The concomitant plot is also the scattergram that we smooth in quantile regression (Carmichael (1978)).

3. Conditional Approach

We have referred before to the complexity of the hypothesis of independence in the context of nonparametric models

e.g. $D(u_1, u_2) = u_1 \cdot u_2$, for all $0 \le u_1, u_2 \le 1$

If we could reduce the dimension of this problem, we might be able to tackle it successfully.

Let

$$D_1(u_1, u_2) = P(U_2 \le u_2 | U_1 = u_1)$$

$$= \int_0^{u_2} d(u_1, t)\, dt \ .$$

Under the hypothesis of independence,

H_3: $D_1(u_1, u_2) = u_2$, $0 \le u_2 \le 1$, for any fixed $0 \le u_1 \le 1$.

Note that to preserve the equivalence between H_1 and H_3 , we have to consider all the values of u_1 . The simplification we have

achieved is that $D_1(u_1, u_2)$ is a density. And, if we look at it as a function of u_1 for fixed u_2 , it is constant. Thus it can be estimated by the autoregressive method and tested to be a "white noise" density. This can be seen as follows:

$$D(u_1, u_2) = \int_0^{u_1}\int_0^{u_2} d(t_1, t_2)\, dt_2\, dt_1$$

$$= \int_0^{u_1} P(U_2 \le u_2 | U_1 = t_1)\, dt_1 \ , \quad \text{as } U_1 \sim U(0,1)$$

$$\text{So } \frac{\partial}{\partial u_1} D(u_1, u_2) = P(U_2 \le u_2 | U_1 = u_1) = D_1(u_1, u_2) \ .$$

Its Fourier coefficients are

$$\varphi_{u_2}(v) = \int_0^1 e^{2\pi i v u_1} D_1(u_1, u_2)\, du_1 \ .$$

We usually normalize so that $\varphi_{u_2}(0) = 1$. As an estimator, we use

$$\tilde{\varphi}_{u_2}(v) = \frac{\sum_{j=1}^{n} e^{2\pi i v (j-\frac{1}{2})/n}\, \tilde{D}_1(\frac{j-1}{n}, u_2)}{\sum_{j=1}^{n} \tilde{D}_1(\frac{j-1}{n}, u_2)}$$

where $$\tilde{D}_1(\frac{j-1}{n}, u_2) = \begin{cases} 1 \ , & \text{if } \frac{R_j - 1}{n} \le u_2 \\ 0 \ , & \text{otherwise} \end{cases}$$

In the estimation of $\tilde{\varphi}_{u_2}(v)$, there are only $n \cdot u_2$ terms that are not equal to zero. For u_2 small, this is a problem. We can estimate $\tilde{\varphi}_{u_2}(v)$ in such a way there are at least $n/2$ terms that are not zero by working with the functions $\bar{D}(u_1, u_2)$ and $\bar{D}_1(u_1, u_2)$ where $\bar{D}(u_1, u_2) = P(U_1 \le u_1, U_2 \ge u_2)$ and $\bar{D}_1(u_1, u_2) = P(U_2 \ge u_2 | U_1 = u_1)$ so that $\bar{D}_1(u_1, u_2) = 1 - D_1(u_1, u_2)$. For values of u_2 less than .05, we would estimate

$$\tilde{\bar{D}}_1\left(\frac{j-1}{n}, u_2\right) = \begin{cases} 1, \text{ if } \frac{R_j - 1}{n} \ge u_2 \\ 0, \text{ otherwise} \end{cases}$$

For each value of u_2, we can compute a set of Fourier coefficients $\{\tilde{\varphi}_{u_2}(v), v = 0, 1, \ldots\}$. If $\frac{k-1}{n} < u_2 < \frac{k}{n}$, these coefficients are constant in u_2. Thus, we consider only the n sets of coefficients obtained for $u_2 = \frac{k}{n}$, $k = 0, 1, \ldots, n-1$. For each set, we compute the autoregressive estimators of $D_1(\cdot, \cdot)$ and use Parzen's CAT criterion to test for constancy: if the order chosen by the CAT criterion is zero, then $D_1(\cdot, \cdot)$ is taken to be constant. We obtain a *vector of orders* determined by the CAT criterion that can be used as a test statistic. It was found empirically that the CAT criterion with sample size taken to be n chose orders different from zero mostly for values of u_2 near 0.5 where only $n/2$ terms contribute to $\tilde{\varphi}_{u_2}(\cdot)$. This suggests modifying the CAT criterion depending on the value of u_2.

If we compute for each u_2, the CAT criterion using as sample size the number of terms that contribute to $\tilde{\varphi}_{u_2}(\cdot)$, we obtain fewer false rejections but the power is considerably decreased. Compared to Spearman test, these new tests don't fare very well.

Table 2

% Correct Decisions for Sample Size 20

ρ	CAT 1	CAT 2	Spearman
0.0	78	98.4	99
0.1	29	1	1
0.2	26	2	3
0.3	35	6	7
0.5	49	15	28
0.6	62	18	48
0.7	70	37	76
0.8	92	46	92
0.9	96	83	100
0.95	100	96	100

For Spearman test, we used the critical values for a two-sided test at $\alpha = 0.01$ (1827 and 2583).

For CAT 1, the cut-off point was $-1-1/n$ with $n = 20$ used as sample size.

For CAT 2, the cut-off point was $-1 - 1/m$ where m was the number of terms contributing to $\tilde{\varphi}_{u_2}(\cdot)$.

4. The Density-Regression Function

Parzen chose the term "density-regression" function for

$$d(u_1, u_2) = \frac{f_{X,Y}\, Q_X(u_1),\ Q_Y(u_2)}{f_X\, Q_X(u_1) \cdot f_Y\, Q_Y(u_2)}$$

It was noted that under the hypothesis of independence, $d(u_1, u_2) \equiv 1$.

We can estimate $d(\cdot, \cdot)$ using Fourier transforms:

$$\hat{d}(u_1, u_2) = \sum_{v_1, v_2 = -\infty}^{\infty} e^{-2\pi i(u_1 v_1 + u_2 v_2)} k_M(v_1, v_2)\, \tilde{\varphi}(v_1, v_2)$$

where $\tilde{\varphi}(v_1, v_2)$ is the characteristic function of the empirical c.d.f. $\tilde{D}(u_1, u_2)$; $k_M(v_1, v_2)$ is a weight function such that the doubly-infinite summation is truncated, e.g., $k_M(v_1, v_2) = g_M(v_1) \cdot g_M(v_2)$, with $g_M(\cdot)$ the Parzen kernel.

For simplicity, we fixed $u_1 = 1/2$ and looked at $\hat{d}(1/2, u_2)$.

In the bivariate normal case, we computed the ratio

$$C = \frac{\max_{.05 \le u_2 \le .95} d(1/2, u_2)}{\min_{.05 \le u_2 \le .95} d(1/2, u_2)}$$

and produced the following table *for different values of the correlation* coefficient ρ .

Table 3

Some Characteristics of the Bivariate Normal

ρ	C	$\varphi(1, -1)$	$\varphi(1, 0)$
.1	1.01	.026	(-.047, .001)
.3	1.14	.089	(-.052, .002)
.5	1.57	.191	(-.060, .003)
.7	3.67	.366	(-.069, .004)
.9	327.71	.705	(-.069, .003)

We also looked at the bivariate characteristic function and found that, as ρ increased, the most important coefficient was $\varphi(1, -1)$ in the sense that $|\varphi(1, -1)|^2 > |\varphi(j, k)|^2$, j and $k \neq 0$.

Based on 50 samples of size 20 , we estimated $\hat{d}(1/2, u_2)$ with $M = 3$ and used as a test statistic

$$C^* = \frac{\max\{\hat{d}(1/2, j/20),\ j = 1, \ldots, 19\}}{\min\{\hat{d}(1/2, j/20),\ j = 1, \ldots, 19\}}$$

Table 4

	$P(C^* > 3.4)$
$\rho = 0.0$	3/50
$\rho = 0.5$	6/50
$\rho = 0.9$	44/50

5. Conclusion

It would seem that the CAT criterion needs to be modified in these contexts because the way it is ordinarily used leads to probability of false rejection much too high.

References

Carmichael, J.-P. (1978). "Techniques of Quantile Regression," Grant Technical Report No. ARO-5, *Statistical Science* Division, State University of New York at Buffalo.

Parzen, E. (1977). "Nonparametric Statistical Data Science: A Unified Approach Based on Density Estimation and Testing for 'White Noise'," Technical Report No. 47, Statistical Science Division, State University of New York at Buffalo.